SYSTEM AND METHOD FOR FULL-TEXT RETRIEVAL AND RECORD MEDIUM WHERE FULL-TEXT RETRIEVING PROGRAM IS RECORDED

Patent number: JP10312395

Publication date: 1998-11-24

Inventor: KITAGAWA RYOKO; SHIRAI TADAHIRO; SUZUKI YOSHIAKI; SUGIYAMA SHINYA; SUGAYA TOMOHIDE

Applicant: TOSHIBA CORP

Classification:

- international: G06F17/30

- european:

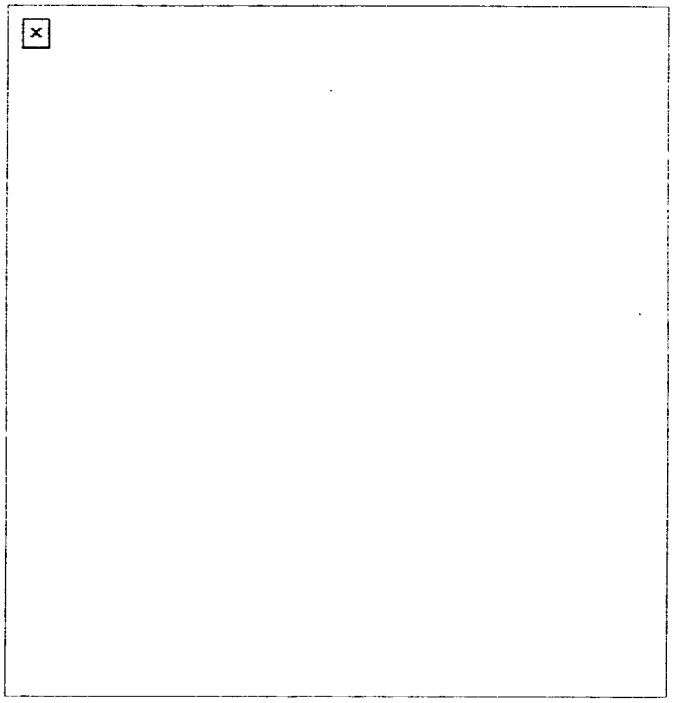
Application number: JP19970324462 19971126

Priority number(s):

Abstract of **JP10312395**

PROBLEM TO BE SOLVED: To improve the retrieval precision by reducing retrieval noise and to reduce the necessary storage capacity of a data base.

SOLUTION: The full-text retrieval system has plural kinds of retrieval information tables 18 to 22 that specify documents which are set to the different numbers of characters on character units and include specified character units, generates plural kinds of character unit groups from an inputted retrieval key word, selects the retrieval information tables by the generated character unit groups, and performs retrieval from the selected retrieval information tables in the generated character units to specify a document including the retrieval key word among respective documents including the retrieved character units. In this case, the number of retrieval information tables is set to a different value by the kinds of characters of a character string including character units set in respective retrieval information tables, and the kind of characters of the retrieval key word is decided to select a retrieval information table to be retrieved according to the character kind and the number of characters of the character string.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平10－312395

(43)公開日　平成10年(1998)11月24日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/40	３７０Ａ
			15/411	３１０
			15/413	３１０Ａ
				３１０Ｂ

審査請求　未請求　請求項の数７　ＯＬ　（全 15 頁）

(21)出願番号　特願平9－324462

(22)出願日　平成９年(1997)11月26日

(31)優先権主張番号　特願平9－54729

(32)優先日　平９(1997)３月10日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　北川　良子
東京都府中市東芝町１番地　株式会社東芝府中工場内

(72)発明者　白井　直裕
東京都府中市東芝町１番地　株式会社東芝府中工場内

(72)発明者　鈴木　善昭
東京都府中市東芝町１番地　株式会社東芝府中工場内

(74)代理人　弁理士　鈴江　武彦　（外６名）

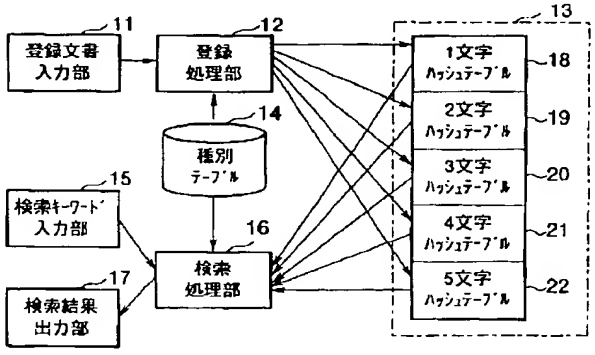
最終頁に続く

(54)【発明の名称】　全文検索システム及び全文検索方法並びに全文検索プログラムを記録した記録媒体

(57)【要約】

【課題】　検索ノイズを低減して検索精度を向上できると共に、データベースの必要記憶容量を低減する。

【解決手段】　それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブル１８～２２を有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から検索キーワードが含まれる文書を特定する全文検索システムにおいて、検索情報テーブルの設定数を、各検索情報テーブルに設定される文字単位を含む文字列の文字種別毎に異なる値に設定し、かつ検索キーワードの文字種別を判定して、この文字種別と文字列の文字数とから検索すべき検索情報テーブルを選択する。



【特許請求の範囲】
【請求項１】　それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを前記生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から前記検索キーワードが含まれる文書を特定する全文検索システムにおいて、
前記検索情報テーブルの設定数は、各検索情報テーブルに設定される文字単位を含む文字列の文字種別毎に異なる値を有し、
前記検索キーワードの文字種別を判定して、この文字種別と文字列の文字数とから前記検索すべき検索情報テーブルを選択することを特徴とする全文検索システム。
【請求項２】　それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを前記生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から前記検索キーワードが含まれる文書を特定する全文検索システムにおいて、
前記各文字単位は、この文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示され、
前記複数種類の検索情報テーブルには、文字数が互いに異なる複数種類の文字単位に対応するハッシュ値が組込まれた共通検索情報テーブルが含まれることを特徴とする全文検索システム。
【請求項３】　それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを前記生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から前記検索キーワードが含まれる文書を特定する全文検索システムにおいて、
前記各文字単位は、この文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示され、
前記複数種類の検索情報テーブルには、前記文字単位を含む文字列の文字種別毎に異なるハッシュ関数を用いて算出されたハッシュ値が設定された検索情報テーブルが含まれ、
前記検索キーワードの文字種別を判定して、文字数と文字種別とから前記検索すべき検索情報テーブルを選択することを特徴とする全文検索システム。
【請求項４】　それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを前記生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から前記検索キーワードが含まれる文書を特定する全文検索方法において、
前記検索情報テーブルの設定数を、各検索情報テーブルに設定される文字単位を含む文字列の文字種別毎に異なる値に設定し、
前記検索キーワードの文字種別を判定して、この文字種別と文字列の文字数とから前記検索すべき検索情報テーブルを選択する全文検索方法。
【請求項５】　それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを前記生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から前記検索キーワードが含まれる文書を特定する全文検索システムにおける全文検索プログラムを記録したコンピュータ読取り可能な記録媒体であって、
前記検索キーワードの設定数を、各検索情報テーブルに設定される文字単位を含む文字列の文字種別毎に異なる値とさせ、
前記検索キーワードの文字種別を判定させ、この文字種別と文字列の文字数とから前記検索すべき検索情報テーブルを選択させることを特徴とする全文検索プログラムを記録したコンピュータ読取り可能な記録媒体。
【請求項６】　それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを前記生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から前記検索キーワードが含まれる文書を特定する全文検索システムにおける全文検索プログラムを記録したコンピュータ読取り可能な記録媒体であって、
前記各文字単位を、この文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示させ、
前記複数種類の件策情報テーブルには、文字数が互いに異なる複数種類の文字単位に対応するハッシュ値が組み込まれた共通検索情報テーブルを含ませることを特徴とする全部検索プログラムを記録したコンピュータ読取り

可能な記録媒体。

【請求項７】　それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを前記生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から前記検索キーワードが含まれる文書を特定する全文検索システムにおける全文検索プログラムを記録したコンピュータ読取り可能な記録媒体であって、

前記各文字単位をこの文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示させ、

前記複数種類の検索情報テーブルには、前記文字単位を含む文字列の文字種別毎に異なるハッシュ関数を用いて算出されたハッシュ値が設定された検索情報テーブルを含ませ、前記検索キーワードの文字種別を判定させ、文字数と文字種別とから前記検索すべき検索情報テーブルを選択させることを特徴とする全文検索プログラムを記録したコンピュータ読取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、大量の文書から必要とする情報が記載された文書を検索する情報検索システムに係わり、特に比較的長い検索キーワードで必要な文書を検索できる全文検索システム、全文検索方法、及び全文検索プログラムを記録した記録媒体に関する。

【０００２】

【従来の技術】一般の情報検索システムのデータベースに新規の文書（文献）を登録する場合は、該当文書に含まれる複数のキーワードをデータベースに登録する。このキーワードは一般に予め決められた単語で構成されている。そして、この情報検索システムを用いて必要な情報が記載された文書（文献）を調べる場合は、必要な情報に関係するキーワードでデータベースを検索すると、このキーワードが登録された文書の文書名，発行所，著者，発行日，所蔵場所等の該当文書を特定する情報が検索結果として出力される。

【０００３】このような情報検索システムにおいては、付与したキーワードでしか検索できなかった。この不都合を解消するために文書中の任意の単語や文字列でデータベースを検索できる全文検索システムが開発されている。

【０００４】図１２は全文検索システムの概略構成図である。この全文検索システムは、大きく分けて、検索キーワードが入力されるキーボード等の検索キーワード入力部１と、実際に検索を実行する検索処理部２と、データベース３と、検索結果を出力するＣＲＴ表示装置や印字装置等で構成された検索結果出力部４とで構成されている。

【０００５】データベース３内には、例えば、１文字検索情報テーブル５，２文字検索情報テーブル６、３文字検索情報テーブル７等の複数の検索情報テーブルが設けられている。

【０００６】１文字検索情報テーブル５内には、図１３（ａ）に示すように、ひらかな、カタカナ、漢字，英字，数字等の各１文字単位毎に、該当１文字単位がこのデータベース３に登録された各文書に含まれるか否かの情報が記憶されている。具体的には、図示するように、該当１文字単位が含まれる文書番号＃に［１］のビットが設定され、該当１文字が含まれない文書番号＃に［０］のビットが設定される。

【０００７】２文字検索情報テーブル６内には、図１３（ｂ）に示すように、前記ひらかな、カタカナ、漢字，英字，数字の２文字の全ての組合わせを示す２文字単位毎に、該当２文字単位が含まれる文書番号＃に［１］のビットが設定され、該当２文字単位が含まれない文書番号＃に［０］のビットが設定される。したがって、２文字検索情報テーブル６に設定されている２文字単位の数は１文字検索情報テーブル５に設定されている１文字からなる１文字単位の数のほぼ２乗値となる。

【０００８】３文字検索情報テーブル７内には、図示しないが、ひらかな、カタカナ、漢字，英字，数字の３文字の全ての組合わせを示す３文字単位毎に、該当３文字単位が含まれる文書番号＃に［１］のビットが設定され、該当１文字が含まれない文書番号＃に［０］のビットが設定される。

【０００９】そして、検索処理部２は図１４に示す流れ図に従って検索キーワード入力部１から入力された検索キーワードに対する検索処理を実行する。流れ図のステップＳ１において、検索キーワード入力部１から一つの検索キーワードが入力されると、Ｓ２において、該当検索キーワードを１文字単位に分解する。例えば、図１５に示すように、検索キーワードが［日本の技術］のように２つの単語と１つの助詞で構成された５文字の文字列からなる場合は、［日］［本］［の］［技］［術］のように１文字からなる５つの１文字単位に分割する。Ｓ３にて、５つの各１文字単位に対してそれぞれ１文字検索情報テーブル５を検索して、各１文字単位が含まれる各文書番号＃を得る。

【００１０】Ｓ４にて、入力された検索キーワードが２文字以上で構成されていた場合は、該当検索キーワードを２文字単位に分解する。例えは、検索キーワードが［日本の技術］の場合は、図１５に示すように、［日本］［本の］［の技］［技術］の４つの２文字単位に分割する（Ｓ５）。そして、４つの各２文字単位に対してそれぞれ２文字検索情報テーブル６を検索して、各２文字単位が含まれる各文書番号＃を得る（Ｓ６）。

【００１１】Ｓ７にて、入力された検索キーワードが３

文字以上の文字列で構成されていた場合は、該当検索キーワードを3文字単位に分解する。例えは、検索キーワードが［日本の技術］の場合は、図15に示すように、［日本の］［本の技］［の技術］の3つの3文字単位に分割する（S8）。そして、3つの各3文字単位に対してそれぞれ3文字検索情報テーブル7を検索して、各3文字単位が含まれる各文書番号#を得る（S9）。

【0012】そして、S10において、上述したS3，S6，S9にて実行された全ての検索結果のアンド（AND）結果を得る。具体的には、全ての文字単位が含まれる文書番号#を抽出する。

【0013】例えば、検索キーワードが［日本の技術］の場合は、1文字単位と2文字単位と3文字単位との合計12個の文字単位が全て含まれる文書番号#を抽出して、この文書番号#をこの検索処理部2における検索結果として検索結果出力部4を介して出力する（S11）。

【0014】このようにして、全文検索システムにおいては、2つ以上の単語を含む比較的長い文字列からなる検索キーワードで該当文字列を含む文書をデータベース3から検索できる。

【0015】このように構成された全文検索システムにおいては、長い文字列が含まれる文書を精度よく検索するためには、1，2，3文字検索情報テーブル5，6，7以外にも4文字，5文字の検索情報テーブルが必要である。

【0016】しかし、検索情報テーブルに設定する文字単位の文字数が増加すると、組合わせ数が飛躍的に増加し、データベース3が必要とする記憶容量が大幅に増大する。

【0017】このような不都合を解消するために、データベース3の各検索情報テーブル5，6，7に登録されている1個又は複数の文字からなる文字単位をハッシュ関数F_i（c_1，c_2，…，c_i）を用いて算出されたハッシュ値hで示す技術が開発されている。

【0018】図16は、文字単位の構成文字数がiである場合における各文字単位をハッシュ値h（＝1，2，3，…，Ni）で示した場合の検索情報テーブルであるi文字ハッシュテーブル8を示す。

【0019】このハッシュ関数F_i（c_1，c_2，…，c_i）は、各文字単位を構成するi個の各文字c_1，c_2，…，c_iの関数で示される。したがって、この関数F_iの式を調整することによって、各文字単位の各文字c_1，c_2，…，c_iの複数種類の組合わせに対して同一ハッシュ値hを設定可能である。

【0020】その結果、i文字数で構成される文字単位の全ての組合わせの数より、ハッシュ値hの数を大幅に減少できる。よって、このハッシュ関数F_i（c_1，c_2，…，c_i）を用いることによって、各検索情報テーブルの記憶容量をある程度低減できる。

【0021】

【発明が解決しようとする課題】しかしながら、上述したハッシュ関数F_i（c_1，c_2，…，c_i）を用いた全文検索システムにおいても、まだ解消すべき次のような課題があった。すなわち、前述したように、文字数の大きい文字列を精度よく検索するには、高い文字数の検索情報テーブル（文字ハッシュテーブル）を設ける必要があるが、データベース3の記憶容量の制約からせいぜい図12に示したように、1，2，3文字検索情報テーブル（文字ハッシュテーブル）までである。

【0022】したがって、図15に示したような5文字からなる文字列「日本の技術」を検索する場合は、5文字の文字列そのもので検索情報テーブルを検索していないので、実際には目的の文字列が含まれない文書が検索されてしまうこともあり、検索精度が低下する。なお、検索結果に混入するこの誤った文書を「検索ノイズ」と称する。

【0023】検索精度が低下するのは、文字種の多い日本語の漢字よりも文字種の少ない英字，数字でよく発生することが知られている。文字種が少ない場合は、検索する文字列の並びが文書中に多数出現する場合が多い。例えば、0～9の数値では、3文字の組合せが$10\times10\times10=10^3$通りしかない。

【0024】しかしJIS規格の第一水準のかな漢字は約3000個存在するので、3文字の組合せが$3000\times3000\times3000=2.7\times10^{10}$通りあるので、それぞれの出現頻度が低くなる。

【0025】例えば、5文字からなる「10000」の数値を検索する場合、各検索情報テーブルに対して次の検索が行われる。

1文字テーブル　1，0，0，0，0
2文字テーブル　10，00，00，00
3文字テーブル　100，000，000

次に、4文字からなる「1000」の数値を検索することを考えると、各検索情報テーブルに対して次の検索が行われる。

【0026】1文字テーブル　1，0，0，0
2文字テーブル　10，00，00
3文字テーブル　100，000

この5文字の文字列「10000」の検索と4文字の文字列「1000」の検索とは、同じ文字単位の検索を複数回行うため、実際には両者ともに次に示す全く同じ検索を行うことになる。

【0027】1文字テーブル　1，0
2文字テーブル　10，00
3文字テーブル　100，000

すなわち、出現頻度の高い「00」の文字単位を何度も検索することになり、たとえ各テーブルにおけるそれぞれの検索結果のANDをとったとしても、同一の検索結果が出力される懸念があり、検索精度が低下する。

【００２８】そのため、文字列「１００００」を含む文書を検索しても、文字列「１０００」を含む文書が同時に検索されてしまう。また、５文字の文字列「１２０００」を検索する場合は、次の文字単位を検索することになる。

【００２９】１文字テーブル　１，２，０，０，０
２文字テーブル　１２，２０，００，００
３文字テーブル　１２０，２００，０００
そのため、次のような文書も誤って検索してしまう。

【００３０】「２０００件の中で１２０件が・・・」
このように、検索する文字列の文字種が少ない場合は、検索する文字列の文字並びが文書中に多数出現する場合が多く、前述した検索ノイズが増加する傾向にある。

【００３１】そこで４文字、５文字を検索するための４文字検索情報テーブル、５文字検索情報テーブルを増設することが考えられるが、データベース３の必要記憶容量が飛躍的に増大し、実用的でないという問題が発生する

【００３２】また、データベース３の必要記憶容量を低減させるために、前述した図１６に示すハッシュ関数F_i（c_1，c_2，…，c_i）を採用したハッシュテーブル８を採用することが考えられる。

【００３３】しかし、ハッシュ関数F_i（c_1，c_2，…，c_i）を使用する場合は、前述したように、必然的に、異なる文字単位が同じハッシュ値ｈを持つ可能性があるため、同じハッシュ値ｈを持つ別の文字単位を含む文書が検索ノイズとして検索される懸念がある。

【００３４】特にハッシュテーブル８中のばらつきに片寄りがある場合は、検索ノイズが増加する。一般的に、このハッシュ関数F_i（c_1，c_2，…，c_i）の設定は難しく、ハッシュ関数F_iの設定の精度によって検索ノイズの発生率が増減する。

【００３５】このように、従来の全文検索システムにおいては、検索する際には、検索キーワードを１文字単位、２文字単位、３文字単位ずつ区切った組合わせで検索するため、４文字以上の文字列の場合、特に数値などの文字種別において、正しい検索結果が得られない懸念がある。

【００３６】また、高い検索精度を得るために文字単位に含まれる文字数が多い多数の検索情報テーブルを組込むことが考えられるが、データベース３の必要記憶容量が大幅に増加する問題があった、さらに、ハッシュ関数F_iを用いてデータベース３の必要記憶容量を減少させる場合は、ハッシュ関数F_iの設定の精度によって検索ノイズの発生率が増加する懸念がある。

【００３７】本発明はこのような事情に鑑みてなされたものであり、検索情報テーブルの設定数を検索すべき文字列の文字種別に応じた値に設定することによって、検索情報テーブルを記憶するデータベースの記憶容量を大幅に増加することなく、検索精度を向上できる全文検索システム及び全文検索方法並びに全文検索プログラムを記録した記録媒体を提供することを目的とする。

【００３８】また、各検索情報テーブルの構成を工夫することによって、検索精度を低下することなく、検索情報テーブルを記憶するデータベースの記憶容量を低減できる全文検索システムを提供することを目的とする。

【００３９】
【課題を解決するための手段】本発明は、それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から検索キーワードが含まれる文書を特定する全文検索システムに適用される。

【００４０】そして、上記課題を解消するために請求項１においては、検索情報テーブルの設定数は、各検索情報テーブルに設定される文字単位を含む文字列の文字種別毎に異なる値を有し、検索キーワードの文字種別を判定して、この文字種別と文字列の文字数とから前記検索すべき検索情報テーブルを選択する。

【００４１】また、請求項２においては、各文字単位は、この文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示され、複数種類の検索情報テーブルには、文字数が互いに異なる複数種類の文字単位に対応するハッシュ値が組込まれた共通検索情報テーブルが含まれる。

【００４２】請求項３においては、各文字単位は、この文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示され、複数種類の検索情報テーブルには、文字単位を含む文字列の文字種別毎に異なるハッシュ関数を用いて算出されたハッシュ値が設定された検索情報テーブルが含まれ、検索キーワードの文字種別を判定して、文字数と文字種別とから記検索すべき検索情報テーブルを選択する。

【００４３】請求項４においては、それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から検索キーワードが含まれる文書を特定する全文検索方法において、検索情報テーブルの設定数を、各検索情報テーブルに設定される文字単位を含む文字列の文字種別毎に異なる値に設定し、検索キーワードの文字種別を判定して、この文字種別と文字列の文字数とから検索すべき検索情報テーブルを選択する。

【００４４】このように構成された全文検索システム及

び全文検索方法においては、検索情報テーブルの設定数は、各検索情報テーブルに設定される文字単位を含む文字列の文字種別毎に異なる値を有している。

【００４５】すなわち、文字種別として、例えば英字，数字、ひらかな、カタカナ、漢字第一水準、漢字第二水準、外字等が存在する。そして、前述したように、検索すべき文字列が英字，数字のみで構成されていた場合は、たとえ構成文字数が異なる文字列であったとしても同一文字列を検索する事態が発生する確率が多いので、検索ノイズが発生する確率が高くなる。

【００４６】一方、英字，数字以外のひらかな、カタカナ、漢字第一水準、漢字第二水準、外字等の文字種別においては、文字列に同一文字が多数含まれることは希であるので、少ない文字数の文字単位の検索情報テーブルのみを用いて検索したとしても検索ノイズの発生確率は少ない。

【００４７】したがって、検索ノイズの発生確率が高い英数字で構成される文字列を分離した文字単位に対する検索情報テーブルに対してのみ、構成文字数の高い検索情報テーブルを設ければ、検索ノイズの発生確率が低下する。

【００４８】なお、英字，数字の文字種は漢字に比較して格段に少ないので、例えば４文字や５文字の検索情報テーブルを設けたとしても、各検索情報テーブルを記憶するデータベースの記憶容量が大幅に増加することはない。

【００４９】また、別の発明においては、各検索情報テーブルに登録される文字単位は、この文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示されている。

【００５０】このハッシュ値は各文字数毎の検索情報テーブル毎に設定されるが、たとえ構成文字数が異なる文字単位であっても、発生確率の小さい文字単位どうしであれば、このハッシュ値を一つの共通検索情報テーブルに組込んだとしても検索ノイズの発生確率は大幅に上昇することはない。

【００５１】よって、検索精度を低下させることなく、データベースの必要記憶容量を節減できる。また、別の発明においては、複数種類の検索情報テーブルには、文字単位を含む文字列の文字種別毎に異なるハッシュ関数を用いて算出されたハッシュ値が設定されている。

【００５２】すなわち、検索対象となる文字列の文字種別は、前述したように、英字、数字、ひらかな、カタカナ、漢字第一水準、漢字第二水準、外字等が存在する。一般に、複数の文字からなる文字単位を例えば文書から無作為に抽出した場合は、検出される文字単位の各文字の各組合わせの発生確率は文字種別に応じてそれぞれ異なる。

【００５３】したがって、発生確率の高い文字単位どうしを同一ハッシュ値に設定されることを防止すると、各文字種別毎にハッシュ関数を設定することによって、検索ノイズの発生確率を低減できる。

【００５４】また、英字，数字は文字種が少ないので、同一組合わせに対してできるだけ複数のハッシュ値が設定されないようにハッシュ関数を設定すればよい。さらに、請求項５乃至請求項７の発明は、それぞれ文字単位の文字数が異なる値に設定され、指定された文字単位が含まれる文書を特定する複数種類の検索情報テーブルを有し、入力された検索キーワードから複数種類の文字単位群を生成し、この生成された各文字単位群毎に検索情報テーブルを選択して、この選択した各検索情報テーブルを生成された各文字単位で検索し、検索された各文字単位が含まれる各文書から前記検索キーワードが含まれる文書を特定する全文検索システムにおける全文検索プログラムを記録したコンピュータ読取り可能な記録媒体である。

【００５５】そして、請求項５における全文検索プログラムは、コンピュータに対して、検索キーワードの設定数を、各検索情報テーブルに設定される文字単位を含む文字列の文字種別毎に異なる値とさせ、検索キーワードの文字種別を判定させ、この文字種別と文字列の文字数とから検索すべき検索情報テーブルを選択させる。

【００５６】また、請求項６における全文検索プログラムは、コンピュータに対して、各文字単位を、この文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示させ、複数種類の件策情報テーブルには、文字数が互いに異なる複数種類の文字単位に対応するハッシュ値が組み込まれた共通検索情報テーブルを含ませる。

【００５７】さらに、請求項７における全文検索プログラムは、コンピュータに対して、各文字単位をこの文字単位を構成する１個又は複数の文字からハッシュ関数を用いて算出されたハッシュ値で示させ、複数種類の検索情報テーブルには、文字単位を含む文字列の文字種別毎に異なるハッシュ関数を用いて算出されたハッシュ値が設定された検索情報テーブルを含ませ、検索キーワードの文字種別を判定させ、文字数と文字種別とから前記検索すべき検索情報テーブルを選択させる。

【００５８】このように構成された全文検索プログラムを記録した記録媒体を用いることによって、上述した機能を有していない従来の全文検索シカテムに対しても、簡単に上述した機能を付加することができる。

【００５９】

【発明の実施の形態】以下本発明の各実施形態を図面を用いて説明する。

（第１実施形態）図１は本発明の第１実施形態の全文検索方法を適用した全文検索システムの概略構成を示すブロック図である。

【００６０】この全文検索システムは、大きく分けて、文章をデータベース１３へ登録するため登録文書入力部

１１及び登録処理部１２と、データベース１３から必要な情報を検索するための検索キーワード入力部１５，検索処理部１６及び検索結果出力部１７と、種別テーブル１４とで構成されている。

【００６１】データベース１３内には、図示するように、複数種類の検索情報テーブルとしての１文字ハッシュテーブル１８、２文字ハッシュテーブル１９、３文字ハッシュテーブル２０、４文字ハッシュテーブル２１及び５文字ハッシュテーブル２２の合計５つハッシュテーブルが設けられている。

【００６２】各ハッシュテーブル１８～２２内には、各ハッシュテーブル１８～２２で指定されて１文字～５文字の各文字単位を構成する文字のすべての組合わせをハッシュ関数F_i（c_1，c_2，…，c_i）を用いて算出されたハッシュ値ｈが設定されており、該当ハッシュ値ｈが得られる文字単位を含む文書番号＃が［１］のビットで登録されている。

【００６３】この合計５つハッシュテーブル１８～２２に登録される１文字～５文字の各組合わせの文字単位を含む文字列の文字種別として、この第１実施形態システムにおいては、

(a)　英字，数字のみ

(b)　ひらがな、カタカナ，漢字第一水準を含む

(c)　漢字第二水準，外字のみ

の３種類に区分している。

【００６４】１文字ハッシュテーブル１８内には、図３（ａ）に示すように、(a)，(b)，(c)の３種類のすべての文字種別を含む４０００文字種の各１文字（１文字単位）毎に、該当１文字単位からハッシュ関数F_1（c_1）を用いて算出されたハッシュ値ｈ（＝１～N_1）が設定されている。

【００６５】２文字ハッシュテーブル１９内には、図３（ｂ）に示すように、(a)，(b)，(c)の３種類のすべての文字種別を含む約４０００文字種の２文字単位毎に、該当２文字単位からハッシュ関数F_2（c_1，c_2）を用いて算出されたハッシュ値ｈ（＝１～N_2）が設定されている。

【００６６】３文字ハッシュテーブル２０内には、図３（ｃ）に示すように、(c)の漢字第二水準，外字を除く、(a)，(b)の２種類の文字種別を含む約３５００文字種の３文字単位毎に、該当３文字単位からハッシュ関数F_3（c_1，c_2，c_3）を用いて算出されたハッシュ値ｈ（＝１～N_3）が設定されている。

【００６７】４文字ハッシュテーブル２１内には、図４（ａ）に示すように、(a)の英字．数字の１種類のみの文字種別を含む約５０文字種の４文字単位毎に、該当４文字単位からハッシュ関数F_4（c_1，c_2，c_3，c_4）を用いて算出されたハッシュ値ｈ（＝１～N_4）が設定されている。

【００６８】５文字ハッシュテーブル２２内には、図４（ｂ）に示すように、(a)の英字，数字の１種類のみの文字種別を含む約５０文字種の５文字単位毎に、該当５文字単位からハッシュ関数F_5（c_1，c_2，c_3，c_4，c_5）を用いて算出されたハッシュ値ｈ（＝１～N_5）が設定されている。

【００６９】前記種別テーブル１４内には、図２に示すように、検索キーワード、登録文字列を構成する前述した(a) (b) (c)の３種類の文字種別毎に、検索対象又は登録対象の各ハッシュテーブル１８～２２が登録されている。

【００７０】具体的には、(a)の英字，数字に対しては全てのハッシュテーブル１８～２２が登録され、(b)のひらかな、カタカナ，漢字第一水準に対しては１，２，３文字のハッシュテーブル１８～２０が登録され、(c)の漢字第二水準，外字に対しては１，２文字のハッシュテーブル１８，１９のみが登録されている。

【００７１】次に、登録処理部１２が行うデータベース１３の各ハッシュテーブル１８～２２に対する登録文書入力部１１から入力された文書の登録処理を図５に示す流れ図を用いて説明する。

【００７２】ステップＲ１において、データベース１３に対して登録すべき文書が存在することを確認すると、該当文書を読取る（Ｒ２）。そして、この文書内に登録すべき文字列が存在すると（Ｒ３）、該当文字列を読込み（Ｒ４）、この入力した文字列の文字数Ｋ、及び該当文字列の文字種別を判断する。具体的には、前述した(a)，(b)，(c)に区分する（Ｒ５）。

【００７３】そして、種別テーブル１４から判別された文字種別に対応する使用ハッシュテーブルを特定する（Ｒ６）。特定された使用ハッシュテーブルの数ｎと、使用ハッシュテーブル名をＮａ(1)，Ｎａ(2)，Ｎａ(3)，…，Ｎａ(n)と設定する（Ｒ７）。

【００７４】以上の準備処理が終了すると、使用ハッシュテーブルを特定するインデックスｉを１に初期化する（Ｒ８）。そして、インデックスｉが使用ハッシュテーブル数ｎ以下で、かつ使用ハッシュテーブル名Ｎａ(i)が文字列の文字数Ｋ以下の場合（Ｒ９）、ｉ文字ハッシュテーブルに対する文書番号＃の登録処理を開始する。

【００７５】すなわち、該当文字列をｉ個の連続文字からなる複数の文字単位に分割して（Ｒ１０）、この各文字単位からハッシュ関数F_i（c_1，c_2，…，c_i）を用いて各ハッシュ値ｈを算出して、各ハッシュテーブル１８～２２のうちのｉ文字ハッシュテーブルの該当ハッシュ値ｈの欄に対して該当文書番号＃を登録する（Ｒ１１）。

【００７６】ｉ文字ハッシュテーブルに対する文書番号＃の登録処理が終了すると、インデックスｉに１を加算して（Ｒ１２）、Ｒ９へ戻り、加算された後のインデックスｉが示すハッシュテーブルに対する文書番号＃の登録処理を開始する。

【００７７】Ｒ９にて、加算後のインデックスｉが使用ハッシュテーブル数ｎを越えると、今回読出した文字列に対する選択された全てのハッシュテーブル１８～２２に対する該当文書番号＃の登録処理が終了したと判断して、Ｒ３へ戻り、先に取込んだ文書から次の文字列の読出を開始する。

【００７８】また、検索処理部１６が行う検索キーワード入力部１５から入力された検索キーワードに対する検索処理を図６に示す流れ図を用いて説明する。検索キーワード入力部１５から検索キーワードが入力されると（Ｑ１）、入力検索キーワードの文字数Ｋ，及び該当文字列の文字種別を判断する。具体的には、前述した(a)，(b)，(c)に区分する（Ｑ２）。そして、種別テーブル１４から判別された文字種別に対応する使用ハッシュテーブルを特定する（Ｑ３）。特定された使用ハッシュテーブルの数ｎと、使用ハッシュテーブルの名Ｎａ(1)，Ｎａ(2)，Ｎａ(3)，…，Ｎａ(n)と設定する（Ｓ４）。

【００７９】以上の準備処理が終了すると、使用ハッシュテーブルを特定するインデックスｉを１に初期化する（Ｑ５）。そして、インデックスｉが使用ハッシュテーブル数ｎ以下で、かつ使用ハッシュテーブル名Ｎａ(i)が文字列の文字数Ｋ以下の場合（Ｑ６）、ｉ文字ハッシュテーブルに対する文書番号＃の検索処理を開始する。

【００８０】すなわち、該当文字列をｉ個の連続文字からなる複数の文字単位に分割して（Ｑ７）、この各文字単位からハッシュ関数F_i（c_1，c_2，…，c_i）を用いて各ハッシュ値ｈを算出して、各ハッシュテーブル１８～２２のうちのｉ文字ハッシュテーブルの該当ハッシュ値ｈに対して設定されている各文書番号＃を抽出（検索）する（Ｑ８）。

【００８１】ｉ文字ハッシュテーブルに対する各文書番号＃の検索処理が終了すると、インデックスｉに１を加算して（Ｑ９）、Ｑ６へ戻り、加算された後のインデックスｉに対応するハッシュテーブルに対する各文書番号＃の検索処理を開始する。

【００８２】Ｑ６にて、加算後のインデックスｉが使用ハッシュテーブル数ｎを越えると、今回入力した検索キーワードに対する選択された全てのハッシュテーブル１８～２２に対する文書番号＃の検索処理が終了したと判断して、Ｑ１０へ進み、ｎ個の各ハッシュテーブル１８～２２で検出された検索結果である全ての文書番号＃のＡＮＤ値を得る。具体的には、全ての文字単位が含まれる文書番号＃を抽出する。そして、この軽策結果を検索結果出力部１７へ表示出力する。

【００８３】このように構成された第１実施形態の全文検索システムにおいては、登録する文字列や検索する検索キーワードの文字種別に応じて、登録したり検索に用いるハッシュテーブルの種別（Ｎａ(1)～Ｎａ(n)）と数ｎとが異なる。

【００８４】具体的には、文字種別(a)の英字，数字のみの場合は、１文字ハッシュテーブル１８から５文字ハッシュテーブル２２までの全てのハッシュテーブル５を使用する。逆に、文字種別(b)の英字，数字，ひらがな、カタカナ，漢字第一水準を含む場合は、１文字ハッシュテーブル１８から３文字ハッシュテーブル２０までの３つのハッシュテーブルを使用する。

【００８５】したがって、４文字ハッシュテーブル２１と５文字ハッシュテーブル２２には文字種別(a)の英字，数字のみの組合せの文字単位に対するハッシュ値ｈのみしか設定されていない。英字，数字のみの組合せ数は文字種別(b)における組合せ数に比較して格段に少ないので、たとえこの４文字ハッシュテーブル２１と５文字ハッシュテーブル２２とをデータベース１３に組込んだとしてもデータベース１３の必要記憶容量が大幅に増加することはない。

【００８６】次に、このように構成された第１実施形態の全文検索システムにおける具体的な登録動作及び検索動作を具体例を用いて説明する。先ず、次の２つの文書を登録する場合を説明する。

【００８７】文書１（＃＝１）　「・・・１２０００件のデータ・・・」

文書２（＃＝２）　「・・・２０００件中の１２０件・・・」

まず文書１（＃＝１）を読む。この場合、「１２０００」と「件のデータ」とに分割する。そして、同じ文字種別の文字を読むと、「１２０００」の部分が(a)の英数字なので、種別テーブル１４で対応するハッシュテーブルを調べる。数字の場合は１～５文字ハッシュテーブル１８～２２が指定されているため、「１２０００」の文字列を次のように１～５文字の文字単位に分割して、それぞれのハッシュ値ｈを計算し、各ハッシュテーブル１８～２２に該当文書番号＃（＝１）を追加登録する。

【００８８】

１文字ハッシュテーブル　１，２，０，０，０

２文字ハッシュテーブル　１２，２０，００，００

３文字ハッシュテーブル　１２０，２００，０００

４文字ハッシュテーブル　１２００，２０００

５文字ハッシュテーブル　１２０００

次の「件のデータ」の文字列は、ひらがな、カタカナ、漢字（第一水準）を含む(b)の文字種別である。この(b)の文字種別に対しては、種別テーブル１４に１～３文字のハッシュテーブル１８，１９，２０が指定されている。よって、同様な手法で各ハッシュテーブル１８，１９，２０に対して該当文書番号＃（＝１）を追加登録する。

【００８９】ここで、別の文字種別の文字列がつながる部分は少ない方の指定ハッシュテーブルに対して登録すると設定しておくと、「００件のデータ」の中の次の文字に対して各ハッシュ値ｈを計算し、各ハッシュテーブ

ルに該当文書番号＃（＝１）を追加登録する。
【００９０】
１文字ハッシュテーブル　件，の，デ，－，タ
２文字ハッシュテーブル　０件，件の，のデ，デー，ータ
３文字ハッシュテーブル　００件，０件の，件のデ，のデー，データ
次に、文書２（＃＝２）を読む。同様に同じ文字種別の文字を読み、次の通りに各ハッシュテーブルに該当文書番号＃（＝２）を追加登録する。
【００９１】１文字ハッシュテーブル　２，０，０，０，件，中，の，１，２，０
２文字ハッシュテーブル　２０，００，００，０件，件中，中の，の１，１２，２０
３文字ハッシュテーブル　２００，０００，００件，０件中，件中の，中の１，の１２，１２０
４文字ハッシュテーブル　２０００
５文字ハッシュテーブル　なし
以上で上記各文書（＃＝１，＃＝２）の各ハッシュテーブルに対する登録が終了する。
【００９２】次に、上述したように各文書（＃＝１，＃＝２）が登録されたデータベース１３を検索キーワード「１２０００」を用いて、この文字列「１２０００」を含む文書を検索する場合を説明する。
【００９３】まず検索キーワード「１２０００」が入力されると、この検索キーワードの数字である文字種別(a)を判定して、種別テーブル１４から、文字種別(a)に対応する１文字から５文字の各ハッシュテーブル１８～２２を特定する。まず、検索キーワード「１２０００」を１文字単位に分割する。
【００９４】１，２，０，０，０
これら５つの各１文字単位について、１文字ハッシュテーブル１８に対して検索を行なうと、文書１、２が検索される。次に検索キーワード「１２０００」を２文字単位に分割する。
【００９５】１２，２０，００，００
これら４つの各２文字単位について、２文字ハッシュテーブル１９に対して検索を行なうと、文書１、２が検索される。次に検索キーワード「１２０００」を３文字単位に分割する。
【００９６】１２０，２００，０００
これら３つの各３文字単位について、３文字ハッシュテーブル２０に対して検索を行なうと、文書１、２が検索される。次に検索キーワード「１２０００」を４文字単位に分割する。
【００９７】１２００，２０００
これら２つの各４文字単位について、４文字ハッシュテーブル２１に対して検索を行なうと、文書１のみが検索される。次に検索キーワード「１２０００」を５文字単位に分割する。
【００９８】１２０００
この１つの５文字単位について、５文字ハッシュテーブル２２に対して検索を行なうと、文書１のみが検索される。
【００９９】最後に、これまでの各検索結果のＡＮＤを取ると文書１のみが残り、この文書１（＃＝１）が最終検索結果として出力される。これにより、正しい文書１（＃＝１）のみが検索され、検索ノイズである文書２（＃＝２）は検索されないので、検索精度が向上する。
【０１００】このように、検索ノイズの発生確率が高い英数字等の文字種別で構成される文字列に対するハッシュテーブルに対してのみ、構成文字数の高い４文字や５文字のハッシュテーブル２１，２２を設ければ、検索ノイズの発生確率が低下する。
【０１０１】なお、英数字の文字種は漢字に比較して格段に少ないので、たとえ４文字や５文字のハッシュテーブル２１，２２を設けたとしても、データベース１３全体の必要記憶容量が大幅に増加することはない。
【０１０２】（第２実施形態）図７は本発明の第２実施形態に係わる全文検索システムの概略構成を示すブロック図である。図１に示す第１実施形態と同一部分には同一符号が付してある。したがって、重複する部分の詳細説明は省略されている。
【０１０３】この第２実施形態の全文検索システムのデータベース１３ａ内には、図１に示す第１実施形態と同一の１文字ハッシュテーブル１８、２文字ハッシュテーブル１９、３文字ハッシュテーブル２０の他に、共通ハッシュテーブル２３が設けられている。
【０１０４】この共通ハッシュテーブル２３内には、図９に示すように、構成する文字数が４文字，５文字，６文字，…，ｎ文字とそれぞれ異なる値を有する各文字単位に対応する各文字組合わせ毎に算出されたハッシュ値ｈが登録されている。
【０１０５】具体的には、図示するように構成文字数ｉ毎に、ハッシュ関数F_iが設定されている。例えば４文字の場合はハッシュ関数F_4（c_1，c_2，c_3，c_4）を用いてハッシュ値ｈを算出する。また、５文字の場合はハッシュ関数F_5（c_1，c_2，c_3，c_4，c_5）を用いてハッシュ値ｈを算出する。さらに、６文字の場合はハッシュ関数F_6（c_1，c_2，c_3，c_4，c_5，c_6）を用いてハッシュ値ｈを算出する。
【０１０６】また、種別テーブル１４ａ内には、検索キーワード及び登録文字列の前述した(a)，(b)，(c)の文字種別毎に利用する各ハッシュテーブルが登録されている。
【０１０７】(a)に示す英字、数字のみの場合、１文字ハッシュテーブル１８、２文字ハッシュテーブル１９、３文字ハッシュテーブル２０及び共通ハッシュテーブル２３が登録されている。なお、文字種別(b)，(c)に対しては図２で示した第１実施形態の種別テーブル１４と

同一のハッシュテーブルが設定されている。

【0108】このような構成の第2実施形態の全文検索システムにおいて、登録処理部12は登録文書入力部11から入力された各文書のそれぞれ登録すべき各文字列を例えば1～N個の文字からなる各単位文字に分離して、種別テーブル14aの指定する各ハッシュテーブルへ該当文章番号#を登録するが、4文字以上の文字単位に対する文書番号#の登録は全て共通ハッシュテーブル23へ一括して登録される。

【0109】検索処理部16は、検索キーワード検索部15から入力された検索キーワードでデータベース13aを検索する場合においても、検索キーワードを1～N個の文字からなる各文字単位に分離して、種別テーブル14aの指定する各ハッシュテーブルを検索するが、4文字以上の文字単位に対する検索はすべて共通ハッシュテーブル23に対して実施する。

【0110】このような、共通ハッシュテーブル23を使用したとしても、目標とする文書を確実に検索できる。また、図9に示したように、ハッシュ値hは各構成文字数毎のハッシュ関数F_4，F_5，F_6，…，F_n毎に個別の値として求まるが、たとえ構成文字数が異なる文字列であっても、4文字単位，5文字単位，6文字単位等の構成文字数が大きいものは登録されている各文書における発生確率が小さい。したがって、この各文字数毎のハッシュ値hを一つの共通ハッシュテーブル23に組込んだとしても誤った文書が検索される検索ノイズの発生確率は大幅に上昇することはない。

【0111】よって、検索精度を低下させることなく、データベース13aの必要記憶容量を節減できる。

（第3実施形態）図10は本発明の第3実施形態に係わる全文検索システムの概略構成を示すブロック図である。図1に示す第1実施形態と同一部分には同一符号が付してある。したがって、重複する部分の詳細説明は省略されている

【0112】この第3実施形態の全文検索システムのデータベース13b内には、1文字ハッシュテーブル18a、2文字ハッシュテーブル19a、3文字ハッシュテーブル20a，4文字ハッシュテーブル21a、5文字ハッシュテーブル22a，6文字ハッシュテーブル24，7文字ハッシュテーブル25の合計7個のハッシュテーブルが設けられている。

【0113】また、種別テーブル14b内には、図11に示すように、検索キーワード及び登録文字列の(a)，(b1)，(b2)，(c)の合計4種類の文字種別毎に利用する各ハッシュテーブル及び採用する各ハッシュ関数が登録されている。

【0114】(a)の文字種別は、第1実施形態と同様に英字と数字のみであり、この文字種別(a)に対して前述した1文字から7文字までの全てのハッシュテーブル18a～25が使用ハッシュテーブルとしてと登録されている。さらに、各ハッシュテーブル18a～25毎に採用するハッシュ関数F_1，F_2，F_3，F_4，F_5，F_6，F_7が登録されている。各ハッシュ関数F_1～F_7は文字単位の構成文字数iが異なるのみのである同一種類のハッシュ関数F_i（c_1，…，c_i）である。

【0115】(b1)の文字種別は、ひらかなとカタカナのみであり、この文字種別(b1)に対して前述した1文字から3文字までの各ハッシュテーブル18a～20aが使用ハッシュテーブルとしてと登録されている。さらに、各ハッシュテーブル18a～20a毎に採用するハッシュ関数G_1，G_2，G_3が登録されている。各ハッシュ関数G_1～G_3は文字単位の構成文字数iが異なるのみの度同一種類のハッシュ関数G_i（c_1，…，c_i）である。

【0116】(b2)の文字種別は、第一水準の漢字のみであり、この文字種別(b2)に対して前述した1文字から3文字までの各ハッシュテーブル18a～20aが使用ハッシュテーブルとしてと登録されている。さらに、各ハッシュテーブル18a～20a毎に採用するハッシュ関数G_1，D_2，D_3が登録されている。各ハッシュ関数D_1～D_3は文字単位の構成文字数iが異なるのみの同一種類のハッシュ関数D_i（c_1，…，c_i）である。

【0117】(c)の文字種別は、第1実施形態と同様に第一水準の漢字と外字のみであり、この文字種別(c)に対して前述した1文字ハッシュテーブル18aと2文字ハッシュテーブル19aが使用ハッシュテーブルとしてと登録されている。さらに、各ハッシュテーブル18a，19aにそれぞれ採用するハッシュ関数E_1，E_2が登録されている。各ハッシュ関数E_1，E_2は文字単位の構成文字数iが異なるのみの同一種類のハッシュ関数E_i（c_1，…，c_i）である。

【0118】このように、検索キーワード及び登録文字列の前述した(a)，(b1)，(b2)，(c)の合計4種類の文字種別毎に異なる種類のハッシュ関数F_i，G_i，D_i，E_iが設定されている。

【0119】登録処理部12aは、登録文書入力部11から入力された各文書の各登録すべき各文字列の文字種別(a)，(b1)，(b2)，(c)を判断して、例えば1～N個の文字からなる各単位文字に分離する。そして、種別テーブル14bの該当文字種別に指定されたハッシュ関数F_i，G_i，D_i，E_iを用いてハッシュ値hを算出し、同じく種別テーブル14bで指定されたハッシュテーブルの該当ハッシュ値hの欄に今回登録しようとする文字列が含まれる文書番号#を追加登録する。

【0120】検索処理部16aは、検索キーワード検索部15から入力された検索キーワードでデータベース13bを検索する場合、検索キーワードを構成する文字の文字種別(a)，(b1)，(b2)，(c)を判断して、例えば1～N個の文字からなる各単位文字に分割する。そして、種別テーブル14bの該当文字種別に指定されたハッシ

ュ関数F_i，G_i，D_i，E_iを用いてハッシュ値hを算出し、同じく種別テーブル14bで指定されたハッシュテーブルの該当ハッシュ値hの欄に登録された文書番号#を読取る。

【0121】このような、検索キーワード、登録する文字列を構成する各文字の文字種別(a)，(b1)．(b2)，(c) 毎に異なるハッシュ関数F_i，G_i，D_i，E_iを用いてハッシュ値hを算出したとしても、目標とする文書を確実に検索できる。

【0122】さらに、この第3実施形形態においては、各文字種別(a)，(b1)．(b2)，(c)毎に異なるハッシュ関数F_i，G_i，D_i，E_iを用いてハッシュ値hを算出している。

【0123】一般に、各文字種別毎に、1文字単位、2文字単位、3文字単位の各文字の組合せの発生状況が異なるので、全ての文字種別(a)，(b1)．(b2)，(c) に亘って同一種別のハッシュ関数を採用してハッシュ値hを算出した場合においては、文字種別によっては、ハッシュテーブル内において、ある特定のハッシュ値hに対して多くの文書番号#が登録されることになる。その結果、検索ノイズの発生確率が上昇したり、ハッシュテーブルを有効に使用できない懸念がある。

【0124】したがって、各文字種別(a)，(b1)．(b2)，(c) 毎に、該当文字種別の組合せの発生状況に対応した最適のハッシュ関数F_i，G_i，D_i，E_iを設定することによって、一つのハッシュ値hに対して多数の文書番号#が登録されることを抑制でき、検索ノイズの発生確率を低下でき、検索精度を向上できる。

【0125】また、4文字から7文字までの各ハッシュテーブル22a，24、25内には、英字及び数字からなる文字単位の組合わせに対するハッシュ値hしか登録されていない。この英字及び数字からなる文字単位の組合わせ数は、漢字の組合わせ数の比較して格段に小さいので、同一の組合わせが同一ハッシュ値hにならないように、この文字種別のハッシュ関数F_4，F_5，F_6，F_7を調整することによって、たとえ検索キーワードに数字が多く含まれる場合であっても、検索ノイズの発生確率を低下でき、検索精度を向上できる。

【0126】なお、本発明は上述した各実施形態のみに限定されるものではない。例えば図1に示した全文検索システムの登録処理部12、検索処理部16の機能をプログラム化し、予めCD－ROMなどの記録媒体に書き込んでおき、このCD－ROMをCD－ROMドライブを搭載した計算機に装填し、計算機がCD－ROMからプログラムをロードすることにより上記実施形態と同様の機能を実現することができる。なお、記録媒体としては、上記CD－ROM以外に、磁気テープ、DVD－ROM、フロッピーディスク、MO）MD．CD－R、メモリカードなどでもよい。

【0127】

【発明の効果】以上説明したように本発明の全文検索システム及び全文検索方法並びに全文検索プログラムを記録した記録媒体においては、検索情報テーブルの設定数を検索すべき文字列の文字種別に応じた値に設定している。したがって、検索情報テーブルを記憶するデータベースの記憶容量を大幅に増加することなく、検索精度を向上できる。

【0128】また、文字数が互いに異なる複数種類の文字列に対応するハッシュ値が組込まれた検索情報テーブルを用いるので、検索精度を低下することなく検索情報テーブルを記憶するデータベースの記憶容量を低減できる。

【0129】さらに、文字列を構成する文字の種別毎に異なるハッシュ関数を用いて算出されたハッシュ値を検索情報テーブルに設定している。したがって、たとえば、発生確率の高い文字単位どうしを同一ハッシュ値に設定されることを防止するように、各文字種別毎にハッシュ関数を設定することによって、検索ノイズの発生確率を低減できる。

【図面の簡単な説明】

【図1】　本発明の第1実施形態に係わる全文検索方法を採用した全文検索システムの概略構成を示すブロック図

【図2】　同全文検索システムに組込まれた種別テーブルの登録内容を示す図

【図3】　同全文検索システムのデータベースに組込まれた各ハッシュテーブルの登録内容を示す図

【図4】　同じく同データベースに組込まれた各ハッシュテーブルの登録内容を示す図

【図5】　同全文検索システムのデータベースに対する文書の登録処理を示す流れ図

【図6】　同全文検索システムのデータベースに対する文書の検索処理を示す流れ図

【図7】　本発明の第2実施形態に係わる全文検索システムの概略構成を示すブロック図

【図8】　同全文検索システムに組込まれた種別テーブルの登録内容を示す図

【図9】　同全文検索システムのデータベースに組込まれた共通ハッシュテーブルの登録内容を示す図

【図10】　本発明の第3実施形態に係わる全文検索システムの概略構成を示すブロック図

【図11】　同全文検索システムに組込まれた種別テーブルの登録内容を示す図

【図12】　従来の全文検索システムの概略構成を示すブロック図

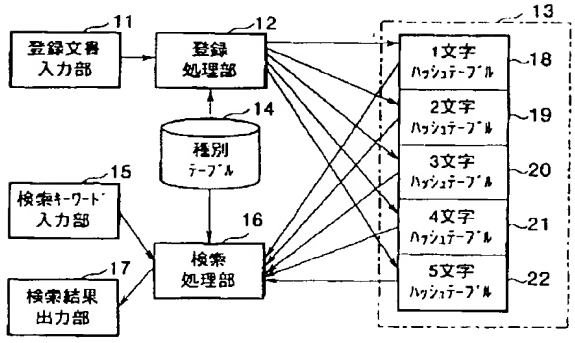
【図13】　同全文検索システムに組込まれた各文字検索情報テーブルの記憶内容を示す図

【図14】　同全文検索システムデータベースに対する文書の検索処理を示す流れ図

【図15】　検索キーワードを文字単位に分割する場合

の分割種別を示す図
【図１６】　一般的な１文字ハッシュテーブルの記憶内容を示す図
【符号の説明】
１１…登録文書入力部
１２，１２ａ…登録処理部
１３，１３ａ，１３ｂ…データベース
１４，１４ａ，１４ｂ…種別テーブル
１５…検索キーワード入力部
１６，１６ａ…検索処理部
１７…検索結果出力部
１８，１８ａ…１文字ハッシュテーブル
１９，１９ａ…２文字ハッシュテーブル
２０，２０ａ…３文字ハッシュテーブル
２１，２１ａ…４文字ハッシュテーブル
２２，２２ａ…５文字ハッシュテーブル
２３…共通ハッシュテーブル
２４…６文字ハッシュテーブル
２５…７文字ハッシュテーブル

【図１】



【図２】

種別テーブル　14

	キーコード、文字列　種別	対応ハッシュテーブル
a	英字、数字のみ	1、2、3、4、5文字ハッシュテーブル
b	ひらがな、カタカナ漢字第一水準、英字、数字を含む	1、2、3文字ハッシュテーブル
c	漢字第二水準、外字のみ	1、2文字ハッシュテーブル

【図３】

1文字ハッシュテーブル　18

(a)

ハッシュ値h ＼ 文書#	1	2	3	4	5	…	…	----------	n
1	0	0	1	0	1				0
2	0	1	0	1	0				1
3	0	0	0	1	1				0
	0	0	0	1	1				1
N1	0	1	0	0	0				0

2文字ハッシュテーブル　19

(b)

ハッシュ値h ＼ 文書#	1	2	3	4	5	…	…	----------	n
1	0	1	0	0	1				1
2	0	1	0	0	1				0
3	1	0	1	0	0				1
N2	0	0	1	0	1				0

3文字ハッシュテーブル　20

(c)

ハッシュ値h ＼ 文書#	1	2	3	4	5	…	…	----------	n
1	0	0	1	0	0				0
2	0	1	0	0	1				0
N3	0	1	0	1	0				1

【図４】

4文字ハッシュテーブル　21

(a)

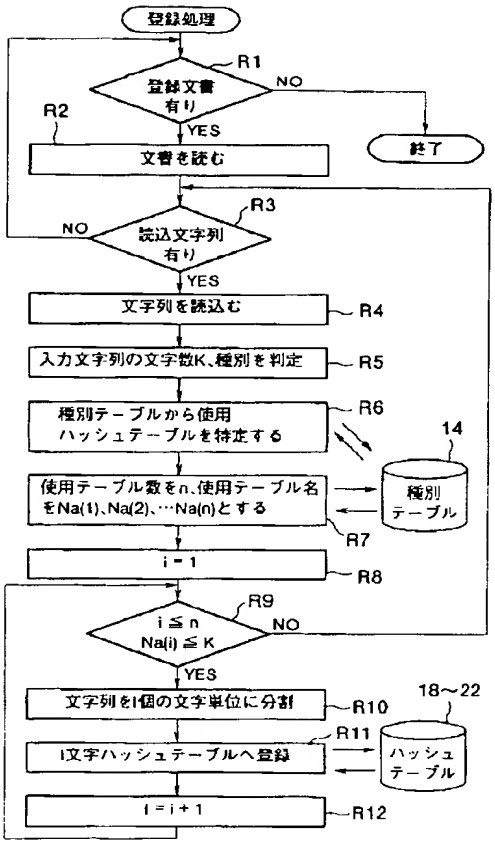
ハッシュ値h ＼ 文書#	1	2	3	4	5	…	…	----------	n
1	0	0	0	0	1				1
2	0	1	0	0	0				0
3	1	0	1	0	0				1
N4	0	0	0	0	1				0

5文字ハッシュテーブル　22

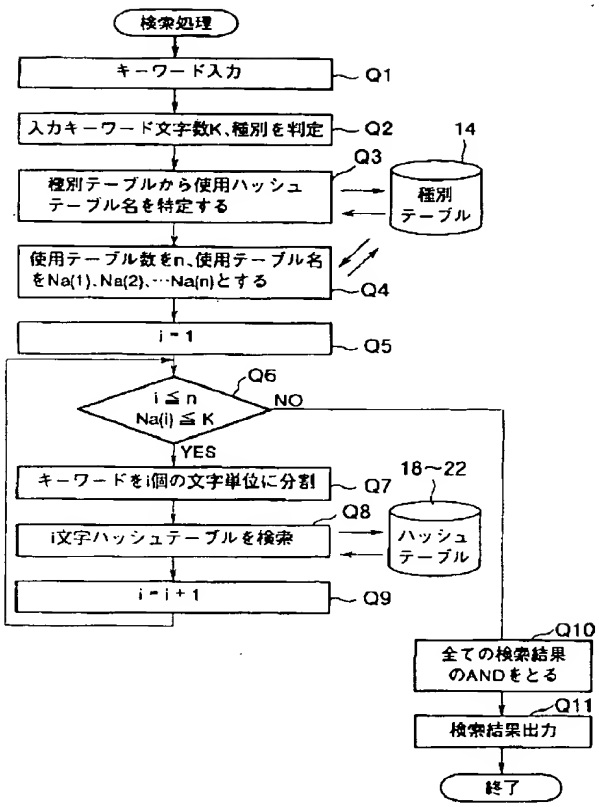
(b)

ハッシュ値h ＼ 文書#	1	2	3	4	5	…	…	----------	n
1	0	0	1	0	0				0
2	0	0	0	0	1				0
N5	0	0	0	1	0				1

【図５】

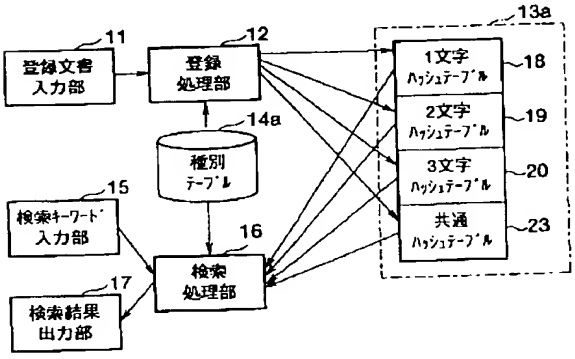


【図６】



【図８】

【図７】



14a

	キーコード、文字列　種別	対応ハッシュテーブル
a	英字、数字のみ	1、2、3文字、共通ハッシュテーブル
b	ひらがな、カタカナ漢字第一水準を含む	1、2、3文字ハッシュテーブル
c	漢字第二水準、外字のみ	1、2文字ハッシュテーブル

【図１２】

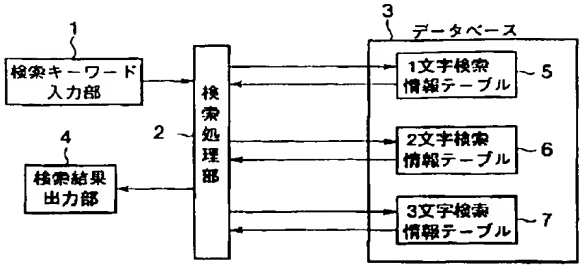
【図１５】

［日本の技術］

1文字単位 日、本、の、技、術

2文字単位 日本、本の、の技、技術、

3文字単位 日本の、本の技、の技術、



【図9】

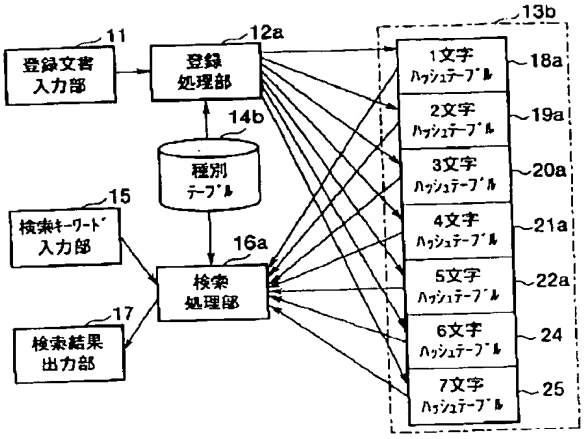
ハッシュ値h ＼ 文書#	1	2	3	4	5	…	…	----------	n
1	0	0	1	0	0				0
2	0	1	0	0	1				0
⋮									
NK	0	1	0	1	0				1

23

ハッシュ値 h
- 4文字の時　F4(C1, C2…, C4)
- 5文字の時　F5(C1, C2…, C5)
- 6文字の時　F6(C1, C2…, C6)
- ⋮
- n文字の時　Fn(C1, C2…, Cn)

Fi；i文字ハッシュ関数 (i = 4, …, n)
Cj；j文字目の文字 (j = 1, …, n)

【図10】



【図11】

14b

	キーワード、文字列　種別	対応ハッシュテーブル	ハッシュ関数
a	英字 数字	1 2 3 4 5 6 7	F1 (C1) F2 (C1,C2) F3 (C1,C2,C3) F4 (C1,…C4) F5 (C1,…C5) F6 (C1,…C6) F7 (C1,…C7)
b1	ひらがな カタカナ	1 2 3	G1 (C1) G2 (C1,C2) G3 (C1,C2,C3)
b2	漢字第一水準	1 2 3	D1 (C1) D2 (C1,C2) D3 (C1,C2,C3)
c	漢字第二水準、外字	1 2	E1 (C1) E2 (C1,C2)

Fi；ハッシュ関数
Gi；ハッシュ関数
Di；ハッシュ関数
Ei；ハッシュ関数
Ci；i番目の文字列

【図13】

(a)

文字単位 ＼ 文書#	1	2	3	4		…….	n
あ	0	1	1	1	0		0
い							
⋮							
東	1	0	1	0	0		1

5

(b)

文字単位 ＼ 文書#	1	2	3	4		…….	n
ああ	0	0	1	0	0		0
あい							
⋮							
東京	1	0	0	0	0		1

6

【図16】

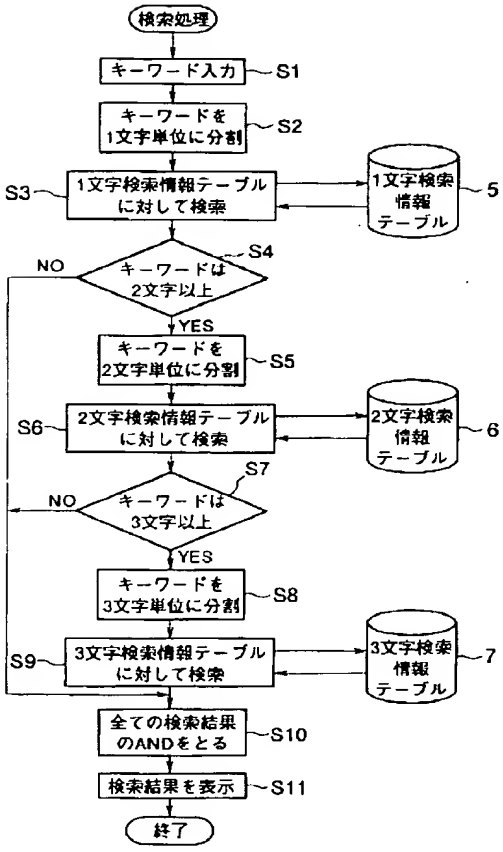
i文字　ハッシュテーブル

ハッシュ値h ＼ 文書#	1	2	3	4	5	…	…	----------	n
1	0	0	1	0	1				0
2	0	1	0	1	0				1
3	0	0	0	1	1				0
	0	0	0	1	1				1
⋮									
Nt	0	1	0	0	0				0

8

h = Fi(C1, C2…Ci)

【図14】



フロントページの続き

(72)発明者　杉山　晋也
東京都府中市東芝町１番地　株式会社東芝
府中工場内

(72)発明者　菅谷　友秀
東京都府中市東芝町１番地　株式会社東芝
府中工場内